

HYBRIDS IN REAL TIME

**Mihaela Baroni, Charles Semple & Mike Steel**

*Department of Mathematics and Statistics  
University of Canterbury  
Private Bag 4800  
Christchurch, New Zealand*

**Report Number:** UCDMS2005/3

APRIL 2005

# HYBRIDS IN REAL TIME

MIHAELA BARONI, CHARLES SEMPLE, MIKE STEEL

**ABSTRACT.** We describe some new and recent results that allow for the analysis and representation of reticulate evolution by non-tree networks. In particular we (1) present a simple result to show how there is always a well-defined ‘tree of life’ even in the presence of reticulation, (2) describe and apply new theory for determining the smallest number of hybridization events required to explain conflicting gene trees, and (3) present a new algorithm to determine whether an arbitrary rooted network can be realised by contemporaneous reticulation events. We illustrate these results with examples.

Corresponding author: Mike Steel

Email: m.steel@math.canterbury.ac.nz

---

*Date:* Submitted 30 March 2005.

2000 *Mathematics Subject Classification.* 05C05, 92D15.

*Key words and phrases.* Directed acyclic graph, reticulate evolution, hybrid species, subtree prune and regraft.

We thank the New Zealand Marsden Fund (UOC-MIS-005) for supporting this research.

## 1. INTRODUCTION

Evolutionary relationships are generally represented by phylogenetic trees, and for certain groups of taxa (e.g. mammals) this model seems well suited. However, for other groups (for example, plants, some fish, and bacteria), processes of reticulate evolution such as the formation of hybrid species, horizontal gene transfer, and other mechanisms (for example endosymbiosis) suggest that evolutionary history would be better described by a network that is more complex than a tree, with some ancestral species arising from the genetic contribution of two (rather than one) lineages.

Although processes of reticulate evolution have long been recognised in biology, techniques for representing and analysing reticulate evolution have tended to be fairly ad-hoc. For example, one might first build a tree and then add some additional edges if these improve the fit of the data, according to a heuristic scheme (as in [15]). In the last few years there has been much new theoretical work by computer scientists and mathematicians [3, 4, 9, 10, 12, 13, 14, 18, 19] with the aim of providing more rigorous approaches to the representation and analysis of reticulate evolution.

In Sections 3 and 4 we provide a brief overview of some of this recent work, and show how it can be applied to set lower bounds on the degree of reticulation required to explain two conflicting phylogenetic trees. We illustrate the application of these results on two trees that describe the evolution of alpine *Ranunculi* in New Zealand. In Section 5, we present a fast algorithm that determines whether or not a hybrid phylogeny can be realised by hybrid events between species that existed at the same time—an obvious biological requirement, though one that can be overlooked in a formal mathematical representation.

## 2. HYBRID PHYLOGENIES

In this section we introduce some terminology that is useful for describing and studying hybrid evolution. Informally, a ‘hybrid phylogeny’ is simply a rooted network in which each arc (directed edge) leads from an ancestral taxon to its immediate descendants. However, unlike a rooted phylogenetic tree, we allow for some (ancestral or extant) taxa to have two (or more) incoming arcs, in other words, that is we regard those taxa as being hybrids, consisting of a genetic composition from both (or all) of the incoming arcs. In this section, we formalize these notions in order to obtain precise results. Furthermore, we describe a tree that underlies any hybrid phylogeny, and provide some background and motivation for the rest of the paper. Throughout, the notation and terminology mostly follows [3, 4].

First we recall some basic graph-theoretic notation. A *directed graph* (or *digraph*)  $D$  is an ordered pair  $(V, A)$  consisting of a non-empty set  $V$  of *vertices* and a subset  $A$  of  $V \times V$  of *arcs*. The *outdegree* (respectively, *indegree*) of a vertex  $v$  of  $D$ , denoted  $d^+(v)$  (respectively,  $d^-(v)$ ) is the number of arcs in  $A$  whose first (respectively, second) component is  $v$ . A *directed cycle* of a digraph is a sequence

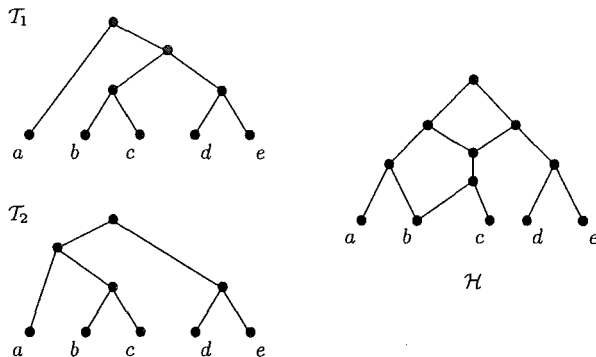


FIGURE 1. A hybrid  $\mathcal{H}$ , and two rooted phylogenetic trees  $\mathcal{T}_1$  and  $\mathcal{T}_2$  displayed by  $\mathcal{H}$ .

$v_0, a_1, v_1, a_2, v_2, \dots, v_{k-1}, a_k, v_k$  of vertices and arcs in which the first and last vertices are equal,  $a_i = (v_{i-1}, v_i)$  for all  $i$ , and, apart from  $v_0$  and  $v_k$ , no vertex or arc appears more than once. A digraph is *acyclic* if it has no directed cycles. An acyclic digraph  $D$  with no underlying parallel edges is *rooted* if there is a distinguished vertex  $\rho$ , called the *root*, with the properties that  $d^-(\rho) = 0$  and there is a directed path from  $\rho$  to every vertex of  $D$ .

Let  $X$  be a finite non-empty set. A *hybrid phylogeny* or *hybrid  $\mathcal{H}$*  on  $X$  is a rooted acyclic digraph in which

- (i)  $X$  is the set of vertices of outdegree zero,
- (ii) the root has outdegree two, and
- (iii) for all vertices  $v$  with  $d^+(v) = 1$ , we have  $d^-(v) \geq 2$ .

Vertices of indegree at least two (called *hybridization vertices*) represent hybridization events and correspond to an exchange of genetic information between hypothetical ancestors. To illustrate, a hybrid phylogeny  $\mathcal{H}$  on  $\{a, b, c, d, e\}$  is shown in Fig 1. Observe that a rooted phylogenetic tree on  $X$  is a particular type of hybrid phylogeny (one that contains no hybridization vertices).

Let  $\mathcal{T}$  be a rooted phylogenetic tree on  $X$  and let  $\mathcal{H}$  be a hybrid phylogeny on  $X'$ , where  $X \subseteq X'$ . Then  $\mathcal{H}$  *displays*  $\mathcal{T}$  if  $\mathcal{T}$  can be obtained from a rooted subtree of  $\mathcal{H}$  by replacing degree-two vertices and their incident edges with a single edge (that is, suppressing degree-two vertices). Extending this to a collection  $\mathcal{P}$  of rooted phylogenetic trees, we say that  $\mathcal{H}$  displays  $\mathcal{P}$  if  $\mathcal{H}$  displays every tree in  $\mathcal{P}$ . For example, in Fig 1, the hybrid phylogeny  $\mathcal{H}$  displays both  $\mathcal{T}_1$  and  $\mathcal{T}_2$ .

**2.1. An underlying tree for a hybrid phylogeny.** Processes of reticulate evolution such as the evolution of hybrid species seem to call into question the very existence of any meaningful concept of a tree of life. However, we now describe a simple mathematical result that formalizes how an underlying ‘tree of life’ always

makes sense (and exists) even in the presence of reticulation. This result is similar in spirit (though different in detail) to results in [1, 9, 14].

Let  $\mathcal{H} = (V, E)$  be a hybrid phylogeny on  $X$  with root vertex  $\rho$ . Let  $V_C$  be the set of vertices of  $\mathcal{H}$  that lie on at least one undirected cycle, and let  $V_T = (V - V_C) \cup \{\rho\} \cup X$ , where  $X$  denotes the set of vertices of outdegree zero. For a vertex  $v$  of  $V$ , let  $c(v)$  denote the set of species  $x$  in  $X$  for which there is a directed path from  $v$  to  $x$  (i.e.  $c(v)$  is the extant species for which  $v$  is an ancestor). A *hierarchy*  $\mathcal{C}$  on  $X$  is a collection of subsets of  $X$ , containing  $X$  and all singleton subsets of  $X$ , and satisfying the property

$$A, B \in \mathcal{C} \Rightarrow A \cap B \in \{\emptyset, A, B\}.$$

Observe that the sets in  $\mathcal{C}$  are *nested*—if they have one or more species in common, then one set is a subset of the other. It is a classical result in phylogenetics that a hierarchy on  $X$  is exactly the set of clusters of a rooted phylogenetic  $X$ -tree. Given a hybrid phylogeny  $\mathcal{H}$ , the following result describes a tree that underlies  $\mathcal{H}$ . Informally speaking, it is the tree obtained by ‘collapsing’ portions of  $\mathcal{H}$  where hybridization has occurred.

**Proposition 2.1.** *Let  $\mathcal{H}$  be a hybrid on  $X$  with vertex set  $V$ . Then the collection  $\mathcal{C} = \{c(v) : v \in V_T\}$  is a hierarchy on  $X$ , in which case there is a rooted phylogenetic  $X$ -tree whose set of clusters is  $\mathcal{C}$ .*

*Proof.* The proof is by contradiction. Suppose that  $\{c(v) : v \in V_T\}$  is not a hierarchy. By definition, there exist vertices  $v_1, v_2 \in V_T$  and elements  $a, b, x \in X$  such that  $x \in c(v_1) \cap c(v_2)$ ,  $a \in c(v_1) - c(v_2)$ , and  $b \in c(v_2) - c(v_1)$ . Since  $c(v_1)$  is not a subset of  $c(v_2)$ , there is no directed path in  $\mathcal{H}$  from  $v_2$  to  $v_1$ . Similarly, there is no directed path from  $v_1$  to  $v_2$ . Since  $x \in c(v_1) \cap c(v_2)$  there is a directed path  $P_1$  from  $v_1$  to  $x$  and a directed path  $P_2$  from  $v_2$  to  $x$ . Let  $v$  be the first vertex that is shared by both  $P_1$  and  $P_2$ . Note that such a vertex exists since  $x$  is a vertex shared by  $P_1$  and  $P_2$ . Since there is no directed path from  $v_1$  to  $v_2$  or  $v_2$  to  $v_1$ , we have that  $v \neq v_1$  and  $v \neq v_2$ . Similarly, there exist directed paths  $Q_i$  from  $\rho$  to  $v_i$  (for  $i = 1, 2$ ) and we can let  $w$  be the last vertex that is shared by  $Q_1$  and  $Q_2$ . Again such a vertex exists since  $\rho$  is shared by both  $Q_1$  and  $Q_2$ . Now if we ignore the direction of the four paths  $P_1$ ,  $P_2$ ,  $Q_1$ , and  $Q_2$  then the path from  $w$  to  $v_1$  (given by  $Q_1$ ) and  $w$  to  $v_2$  (given by  $Q_2$ ) and from  $v_1$  to  $v$  (given by  $P_1$ ) and from  $v_2$  to  $v$  (given by  $P_2$ ) constitutes an undirected cycle in  $\mathcal{H}$ , contradicting the assumption that  $v_1, v_2 \in V_T$ .  $\square$

For the hybrid phylogeny  $\mathcal{H}$  shown in Fig. 1, the above construction yields the rooted phylogenetic tree that has just one cluster other than  $X$  and the singleton subsets of  $X$ , namely  $\{d, e\}$ .

**2.2. Real-time hybrids.** Maddison [17] (see also [18]) pointed out an important biological requirement of hybrid phylogenies. Namely, although a hybrid phylogeny might display two trees, there may be no process of hybridization between contemporaneous taxa (either past or present) that can realise this hybrid phylogeny. Nevertheless, by allowing for additional (unsampled, or perhaps extinct) taxa one

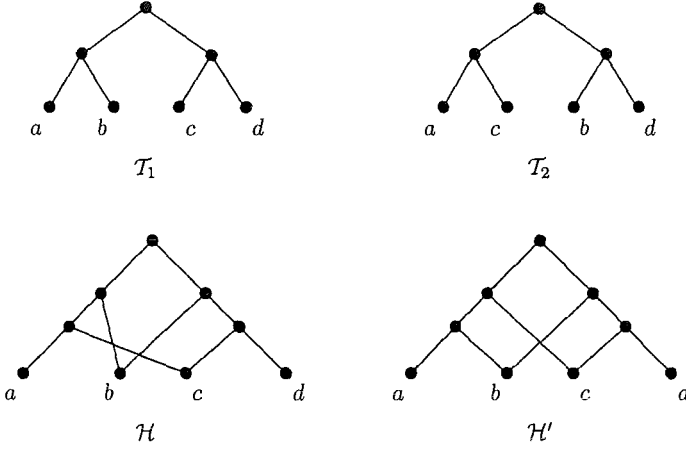


FIGURE 2. Two rooted phylogenetic trees  $T_1$  and  $T_2$  and two hybrid phylogenies  $\mathcal{H}$  and  $\mathcal{H}'$  that display  $T_1$  and  $T_2$ .

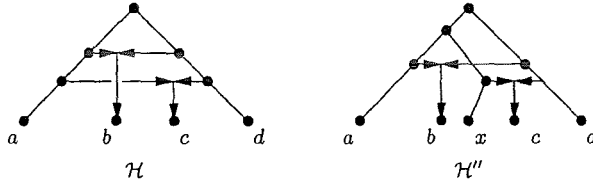


FIGURE 3. Two hybrid phylogenies that explain the real-time evolutionary histories of  $T_1$  and  $T_2$  in Fig. 2.

can resolve this issue without introducing any additional hybridizations. Essentially the role of such an additional taxa is to ‘carry’ a gene (or combination of genes) from the past into some time when it can be inserted into the new hybrid species. Whether these taxa really are (or were) present is another question, but if we are concerned with just placing lower bounds on the degree of hybridization then we can (conservatively) allow them.

To illustrate this point, consider Fig. 2. Both hybrid phylogenies  $\mathcal{H}$  and  $\mathcal{H}'$  display  $T_1$  and  $T_2$  using two hybridization vertices. However, while  $\mathcal{H}$  has a ‘real-time’ realization (see Fig 3),  $\mathcal{H}'$  has no such realization. Nevertheless, by allowing another species  $x$  that may be either extinct or not yet sampled, one can provide such a realization to  $\mathcal{H}'$ . This realization is shown as  $\mathcal{H}''$  in Fig. 3.

In Section 5 we present an algorithm for determining whether a given hybrid phylogeny has a ‘real-time’ realization, or whether additional taxa (as in  $\mathcal{H}''$  in Fig. 3) might be required.

**2.3. Finding the minimal degree of hybridization.** A topical problem in the study of the evolution of a set of species whose past includes reticulation is the following: given an initial set of data that correctly represents the tree-like evolution of different parts of various species genomes, what is the smallest number of reticulation events required that simultaneously explains the entire set of data? This smallest number sets a lower bound on the degree of reticulation that has occurred in the evolution of the species under consideration. If this initial set of data is a collection of rooted phylogenetic trees, this problem can be interpreted within the framework of hybrid phylogenies as follows.

For a hybrid phylogeny  $\mathcal{H}$  with vertex set  $V$  and root  $\rho$ , set

$$h(\mathcal{H}) = \sum_{v \in V; v \neq \rho} (d^-(v) - 1).$$

Observe that  $h(\mathcal{H}) \geq 0$ , and  $h(\mathcal{H}) = 0$  precisely if  $\mathcal{H}$  is a rooted phylogenetic tree. Extending this definition, the *hybrid number* of a collection  $\mathcal{P}$  of rooted phylogenetic trees is

$$h(\mathcal{P}) = \min\{h(\mathcal{H}) : \mathcal{H} \text{ is a hybrid that displays } \mathcal{P}\}.$$

The value  $h(\mathcal{P})$  represents the smallest number of hybridization events that are required to explain  $\mathcal{P}$ . Bordewich and Semple [7] showed that computing this number is NP-hard even in the simplest case that  $\mathcal{P}$  consists of just two rooted binary phylogenetic trees on the same leaf sets. However, despite this negative result, there are some attractive and useful positive results that have recently been described for computing and bounding  $h(\mathcal{P})$  in this setting where  $\mathcal{P}$  consists of two rooted binary phylogenetic trees (on the same leaf sets), and we describe these in Section 3.

### 3. THE MINIMUM NUMBER OF HYBRID EVENTS REQUIRED FOR TWO TREES

Let  $\mathcal{T}$  and  $\mathcal{T}'$  be two rooted binary phylogenetic  $X$ -trees. We will write  $h(\mathcal{T}, \mathcal{T}')$  to denote  $h(\mathcal{P})$  for  $\mathcal{P} = \{\mathcal{T}, \mathcal{T}'\}$ .

The first result we describe shows how one can simplify the calculation of  $h(\mathcal{T}, \mathcal{T}')$  when one or more clusters are shared by both  $\mathcal{T}$  and  $\mathcal{T}'$ . More precisely, suppose that  $A \subset X$  is a cluster of both  $\mathcal{T}$  and  $\mathcal{T}'$  (that is, there is a vertex of each tree that has  $A$  as its set of descendants in  $X$ ). Let  $\mathcal{T}|A$  and  $\mathcal{T}'|A$  denote the subtree of  $\mathcal{T}$  and  $\mathcal{T}'$  (respectively) that have leaf set  $A$ , and let  $\mathcal{T}_a$  and  $\mathcal{T}'_a$  be the rooted trees obtained from  $\mathcal{T}$  and  $\mathcal{T}'$  (respectively) by replacing the subtree having leaf set  $A$  with a new leaf  $a$ .

**Theorem 3.1.** *Let  $\mathcal{T}$  and  $\mathcal{T}'$  be two rooted binary phylogenetic  $X$ -trees. Suppose that  $A \subset X$  is a cluster of both  $\mathcal{T}$  and  $\mathcal{T}'$ . Then*

$$h(\mathcal{T}, \mathcal{T}') = h(\mathcal{T}|A, \mathcal{T}'|A) + h(\mathcal{T}_a, \mathcal{T}'_a).$$

The proof of Theorem 3.1 is given in [3], and the result is typical of other relationships that can be established by exploiting a description of  $h(\mathcal{T}, \mathcal{T}')$  in terms of what has recently been called a “maximum-good-agreement-forest” for

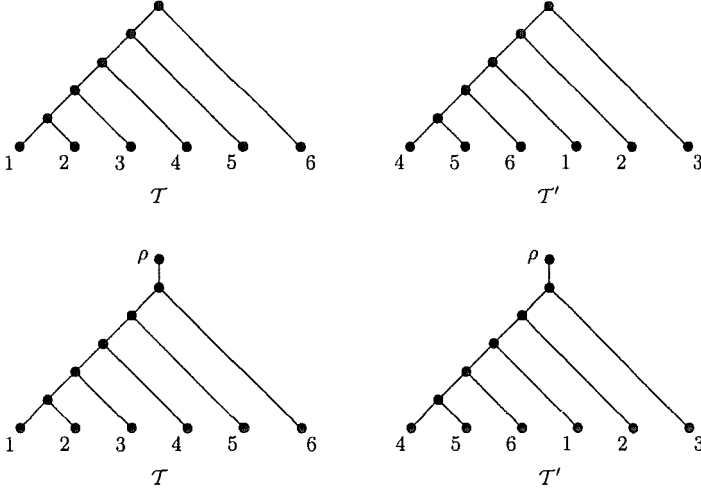


FIGURE 4. Two rooted binary phylogenetic trees  $T$  and  $T'$  without (above) and with (below) their root labelled  $\rho$ .

the pair  $T$  and  $T'$ . We describe this connection now, and provide an application in the next section to show how these results can be used in practice.

To make the interpretation work, we regard the root of both  $T$  and  $T'$  as a vertex  $\rho$  that is adjoined to the original root by a new edge. Furthermore, we view  $\rho$  as part of the label sets of both  $T$  and  $T'$ ; that is, we view the label sets of  $T$  and  $T'$  as  $X \cup \{\rho\}$ . For example, consider the two rooted binary phylogenetic trees  $T$  and  $T'$  shown in the top part of Fig. 4. For the purposes of the interpretation, we view  $T$  and  $T'$  as shown in the bottom part of Fig. 4.

An *agreement forest* for  $T$  and  $T'$  is a collection  $\{\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_k\}$ , where  $\mathcal{T}_\rho$  is a rooted tree whose label set  $\mathcal{L}_\rho$  includes  $\rho$  and  $\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_k$  are rooted binary phylogenetic trees with label sets  $\mathcal{L}_1, \mathcal{L}_2, \dots, \mathcal{L}_k$  such that the following properties are satisfied:

- (i) The label sets  $\mathcal{L}_\rho, \mathcal{L}_1, \mathcal{L}_2, \dots, \mathcal{L}_k$  partition  $X \cup \{\rho\}$ .
- (ii) For all  $i \in \{\rho, 1, 2, \dots, k\}$ ,  $\mathcal{T}_i \cong T|_{\mathcal{L}_i} \cong T'|_{\mathcal{L}_i}$ .
- (iii) The trees in  $\{\mathcal{T}(\mathcal{L}_i) : i \in \{\rho, 1, 2, \dots, k\}\}$  and  $\{\mathcal{T}'(\mathcal{L}_i) : i \in \{\rho, 1, 2, \dots, k\}\}$  are vertex disjoint rooted subtrees of  $T$  and  $T'$ , respectively.

More informally,  $\mathcal{F}$  is an agreement forest for  $T$  and  $T'$  if, up to suppressing degree-two vertices,  $\mathcal{F}$  can be obtained from each of  $T$  and  $T'$  by deleting  $|\mathcal{F}| - 1$  edges. As an example, the two forests  $\mathcal{F}_1$  and  $\mathcal{F}_2$  shown in Fig. 5 are both agreement forests for the two trees  $T$  and  $T'$  shown in Fig. 4.

It has recently been shown ([6]) that for any two rooted binary phylogenetic trees  $T$  and  $T'$  on the same leaf set the smallest value of  $k$  of any agreement forest



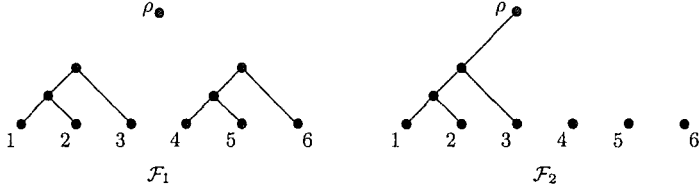


FIGURE 5. Two agreement forests for the two rooted binary phylogenetic trees shown in Fig. 4.

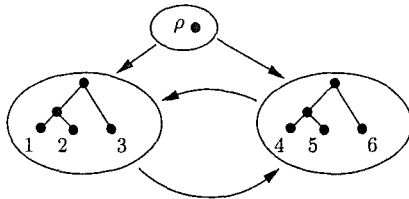
for  $T$  and  $T'$  equals the *rooted subtree prune and regraft distance* between  $T$  and  $T'$ . Denoted  $d_{\text{rSPR}}(T, T')$ , this distance is the minimum number of rooted subtree prune and regraft operations required to transform  $T$  into  $T'$ . It is tempting to conjecture that  $d_{\text{rSPR}}(T, T')$  and  $h(T, T')$  are identical, and indeed the former takes the value 1 if and only if the latter does. However,  $d_{\text{rSPR}}(T, T')$  is only a lower bound for  $h(T, T')$ , and one can construct pairs of trees  $T$  and  $T'$  on  $n$  species such that  $d_{\text{rSPR}}(T, T') = 2$  yet  $h(T, T') > \frac{n}{2} - 1$  [5].

An agreement forest for  $T$  and  $T'$  is a *maximum-agreement forest* if, amongst all agreement forests for  $T$  and  $T'$ , it has the smallest number of components. To continue the previous example, it is straightforward to check that the forest  $\mathcal{F}_1$  in Fig. 5 is a maximum-agreement forest for the two trees  $T$  and  $T'$  in Fig. 4. Thus the rooted subtree prune and regraft distance between these two trees is 2. For the interpretation of  $h(T, T')$  in terms of agreement forest, we need one further definition.

Let  $\mathcal{F} = \{T_\rho, T_1, T_2, \dots, T_k\}$  be an agreement forest for  $T$  and  $T'$ . Let  $G_{\mathcal{F}}$  be the directed graph whose vertices represent the trees in  $\mathcal{F}$  and for which  $(T_i, T_j)$  is a directed edge from the vertex representing  $T_i$  to the vertex representing  $T_j$  precisely if  $i \neq j$  and either

- (I) the root of the subtree  $T(\mathcal{L}_i)$  in  $T$  is an ancestor of the root of the subtree  $T(\mathcal{L}_j)$  in  $T$ , or
- (II) the root of the subtree  $T'(\mathcal{L}_i)$  in  $T'$  is an ancestor of the root of the subtree  $T'(\mathcal{L}_j)$  in  $T'$ .

Since  $\mathcal{F}$  is an agreement forest, the roots of the subtrees  $T(\mathcal{L}_i)$  and  $T(\mathcal{L}_j)$ , and the roots of the subtrees  $T'(\mathcal{L}_i)$  and  $T'(\mathcal{L}_j)$  are not the same. We call  $\mathcal{F}$  a *good-agreement forest* if  $G_{\mathcal{F}}$  is acyclic; that is, if  $G_{\mathcal{F}}$  has no directed cycles. Furthermore, if over all good-agreement forests for  $T$  and  $T'$ ,  $\mathcal{F}$  contains the smallest number of components, then  $\mathcal{F}$  is a *maximum-good-agreement forest* for  $T$  and  $T'$ , in which case we denote this value of  $k$  by  $m_g(T, T')$ . Observe that  $m_g(T, T') = 0$  if and only if, up to isomorphism,  $T$  and  $T'$  are identical. The forest  $\mathcal{F}_2$  in Fig. 5 is a good-agreement forest for the two trees  $T$  and  $T'$  in Fig. 4. Indeed, this forest is maximum-good-agreement forest for  $T$  and  $T'$ . To see that  $\mathcal{F}_1$  is not a good-agreement forest for  $T$  and  $T'$ , observe that  $G_{\mathcal{F}_1}$  contains a directed cycle (see Fig. 6, where the vertices are drawn as large circles enclosing the trees they represent).

FIGURE 6. The graph  $G_{\mathcal{F}_1}$ .

The interpretation of the hybrid number of two rooted binary phylogenetic trees on the same label sets in terms of agreement forests is stated in following theorem which is established in [5].

**Theorem 3.2.** *Let  $T$  and  $T'$  be two rooted binary phylogenetic  $X$ -trees. Then*

$$h(T, T') = m_g(T, T').$$

For example, it follows from Theorem 3.2 that the value of  $h(T, T')$  for the two trees in Fig. 4 is 3.

#### 4. APPLICATION

In this section, we apply the theory of Section 3 to two phylogenetic trees on 46 sequences of alpine Rununculi of New Zealand, reported in [16]. The first tree was constructed from nuclear ITS sequences, while the second was constructed from chloroplast ( $J_{SA}$ ) sequences (for details see [16]). The two trees showed considerable agreement, however there was also a fair degree of incompatibility. One possible explanation for this incompatibility is the occurrence of hybrid evolution, whereby the nuclear ITS sequence has a different history to the chloroplast ( $J_{SA}$ ) sequences. Of course, there may be other sources of phylogenetic error (sampling effects, model mis-specification, lineage sorting) that could cause the two trees to conflict, even in the absence of any hybrid evolution. Nevertheless, we can still ask the following question: Assuming the two trees correctly describe the history of the two genes, and their incongruence is due to hybrid evolution, what is the smallest number of hybrid events required to explain this? The study is complicated slightly by the fact that neither tree is binary. In this case, we took a conservative approach and allowed non-binary subtrees to be resolved in any way that helped minimize the required number of hybridization events. Also, for the sake of illustration in this paper, we will restrict attention to a subgroup (“Group P”) of the sequences consisting of 20 sequences. This group is a candidate for reticulate evolution, since the  $F_1$  progeny are known to be fertile [8]. The two trees for these 20 sequences are shown in Fig. 7, with  $\mathcal{T}_1$  the nuclear, and  $\mathcal{T}_2$  the chloroplast tree.

For  $\mathcal{T}_1$  and  $\mathcal{T}_2$ , one can identify five clusters (denoted  $l_1$  to  $l_5$  in Fig. 7) shared by these two trees; this allows us to apply Theorem 3.1. In this way we reduce the problem from comparing two 20-taxon trees to one of comparing two 5-taxon trees

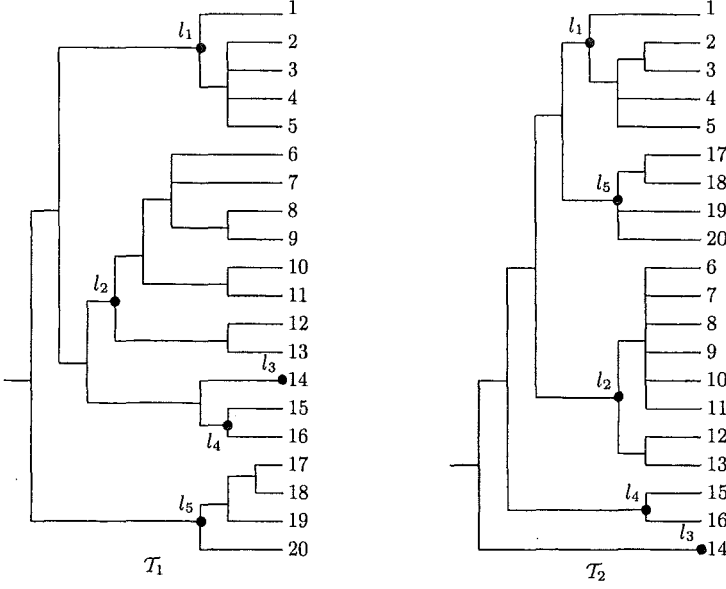


FIGURE 7. The tree  $\mathcal{T}_1$  for nuclear ITS sequences and  $\mathcal{T}_2$  for chloroplast  $J_{SA}$  sequences from [16] restricted to Group I.

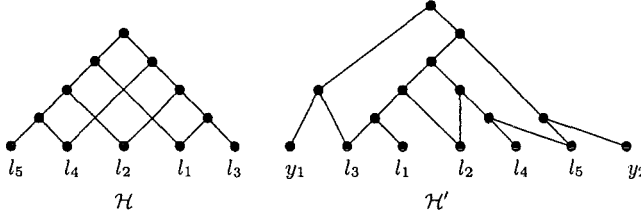


FIGURE 8. Two hybrid phylogenies that display  $\mathcal{T}_1$  and  $\mathcal{T}_2$ , and requiring three hybridization events (the fewest possible for these two trees).

(each having leaf set  $l_1, \dots, l_5$ ), together with the trees on the shared clusters (in fact these latter trees do not contribute to the  $h$  score, since all these pairs of cluster subtrees are compatible). Now using Theorem 3.2, one can show that  $h(\mathcal{T}_1, \mathcal{T}_2) = 3$ . Fig. 8 shows one hybrid phylogeny ( $\mathcal{H}_1^+$ ) that displays the five clusters shared by  $\mathcal{T}_1$  and  $\mathcal{T}_2$  with three hybrid events. Similarly, for the full set of 46 sequences it can be shown (by hand) that the  $h$  value lies between 7 and 12 ([3]). Thus, assuming the trees are correct we require at least 3 hybrid events to describe the evolution of the Group I sequences, and at least 7 hybrid events to describe the evolution of the entire group of 46 sequences. We should stress that this analysis is to illustrate the techniques, rather than to formally show that there has been this degree of hybrid evolution in the taxa described—as we mentioned there are other reasons why trees

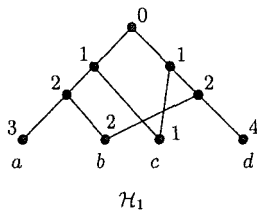


FIGURE 9. A temporal labelling of a hybrid.

may disagree, and these need to be considered (these other processes often leave different statistical signatures from hybridization, see [11, 14]).

It can be checked that the hybrid phylogeny  $\mathcal{H}$  shown in Fig. 8 has no real-time realization (in the sense described in 2.2) however the hybrid phylogeny  $\mathcal{H}'$  in Fig. 8 allows a ‘real-time’ hybrid evolution scenario, with just two extra taxa  $y_1$  and  $y_2$ . Although the analysis of deciding a real time realization could be resolved for this small-scale example by an ad-hoc case analysis, it is clear that such a task could be complicated for a large and complex hybrid phylogeny. In Section 5, we present an algorithm to determine whether an arbitrary hybrid phylogeny can be realised by hybrid evolution between contemporaneous ancestral taxa.

## 5. AN ALGORITHM FOR ‘REAL-TIME’ HYBRIDS

The concept of a ‘real-time hybrid’ has been briefly and informally mentioned already; now we formalize this notion, and provide an algorithm to determine whether an arbitrary hybrid phylogeny can be realized in this way. Some of the more technical parts of this section have been moved to an appendix to assist readability.

Let  $\mathcal{H}$  be a hybrid phylogeny with vertex set  $V$  and arc set  $A$ . We say that  $\mathcal{H}$  has a *temporal representation* if there exists a map  $f : V \rightarrow \mathbb{N}$  with the following properties:

- (i) If  $v$  is a vertex of  $\mathcal{H}$  with  $d^-(v) = 1$ , then  $f(u) < f(v)$  for the (only one) immediate ancestor  $u$  of  $v$ .
- (ii) If  $v$  is a vertex of  $\mathcal{H}$  with  $d^-(v) \geq 2$ , then  $f(u) = f(v)$  for all immediate ancestors  $u$  of  $v$ .

Such a map is called a *temporal labelling* of  $\mathcal{H}$ . To illustrate, a temporal labelling of a hybrid is shown in Fig. 9. All rooted phylogenetic trees have a temporal representation. However, not all hybrid phylogenies have such a representation. For example, the hybrid shown in Fig. 10(a) has no temporal representation.

The main result of this section (Theorem 5.1) is to characterize exactly when an arbitrary hybrid phylogeny has a temporal representation. To this end, we next describe a particular digraph associated with a fixed hybrid  $\mathcal{H}$ . Let  $V$  and  $A$  be the vertex and arc sets of  $\mathcal{H}$ , respectively. We begin by partitioning  $A$  and describing

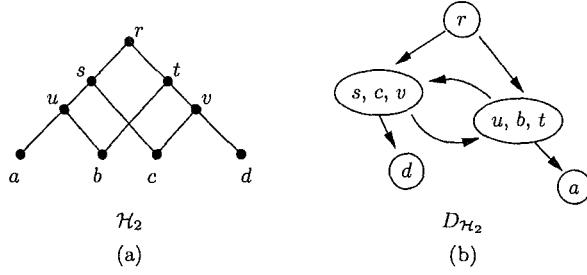


FIGURE 10. (a) A hybrid  $\mathcal{H}_2$  with no temporal representation and (b) its associated digraph  $D_{\mathcal{H}_2}$ .

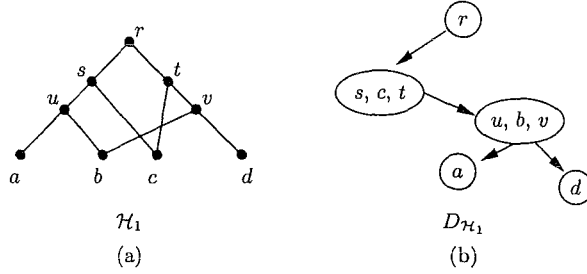


FIGURE 11. (a) A hybrid  $\mathcal{H}_1$  and (b) its associated digraph  $D_{\mathcal{H}_1}$ .

an equivalence relation on  $V$ . Let

$$A_T = \{(u, v) \in A : d^-(v) = 1\}$$

and

$$A_R = \{(u, v) \in A : d^-(v) \geq 2\}.$$

Any arc in  $A_T$  is called a *tree arc* and any arc in  $A_R$  is called a *recombination arc*. Ignoring the direction of the arcs of  $\mathcal{H}$ , an equivalence relation on  $V$  is now defined by setting

$$[v] = \{v\} \cup \{u \in V : \text{there is a path of recombination arcs from } u \text{ to } v \text{ in } \mathcal{H}\}.$$

Observe that if  $v$  is not incident with a recombination arc, then  $[v] = \{v\}$ . Set

$$[V] = \{[v] : v \in V\}.$$

We describe our associated digraph  $D_{\mathcal{H}}$  as follows. The vertex set of  $D_{\mathcal{H}}$  is  $[V]$ , and  $[u]$  and  $[v]$  are joined by an arc  $([u], [v])$  if there exists  $a \in [u]$  and  $b \in [v]$  such that  $(a, b)$  is a tree arc in  $A$ . It is easily seen that  $D_{\mathcal{H}}$  is connected. To illustrate, consider Fig. 11. In Fig. 11(a), we have extended the labelling of the vertices of outdegree zero to all of the vertices of  $\mathcal{H}_1$ , where  $\mathcal{H}_1$  is the hybrid shown in Fig. 9. The digraph  $D_{\mathcal{H}_1}$  is shown in Fig. 11(b). Here

$$[V] = \{\{r\}, \{s, c, t\}, \{u, b, v\}, \{a\}, \{d\}\}.$$

Let  $\mathcal{H}$  be a hybrid and suppose that  $f : V \rightarrow \mathbb{N}$  is a temporal labelling of  $\mathcal{H}$ . Let  $\bar{f}$  be the map from  $[V]$  to  $\mathbb{N}$  that is defined by setting  $\bar{f}([v]) = f(v)$  for all  $v \in V$ . To see that this map is well-defined, first observe that if  $[u] = [v]$ , then there is an (undirected) path from  $u$  to  $v$  consisting of recombination arcs. Since the end vertices of any arc on this path are assigned the same natural number under  $f$ , it follows that all vertices in this path are assigned the same natural number under  $f$ . Hence, for all  $w, w' \in [v]$ , we have  $f(w) = f(w')$ . Thus  $\bar{f}$  is well-defined. Moreover, as  $f$  is a temporal labelling of  $\mathcal{H}$ , there is no  $u$  and  $v$  in the same equivalence class such that  $(u, v)$  is a tree arc.

The following result provides a concise characterization for when a hybrid phylogeny has a temporal representation; its proof is given in the Appendix.

**Theorem 5.1.** *A hybrid  $\mathcal{H}$  has a temporal representation if and only if  $D_{\mathcal{H}}$  is acyclic.*

Theorem 5.1 is the basis for a polynomial-time algorithm (TEMPREP) for determining whether or not a hybrid has a temporal representation and, if so, providing a temporal labelling.

**Algorithm:** TEMPREP( $\mathcal{H}$ )

**Input:** A hybrid phylogeny  $\mathcal{H}$  with vertex set  $V$ .

**Output:** A temporal labelling  $f$  of  $\mathcal{H}$  or the statement  $\mathcal{H}$  has no temporal representation.

1. Construct  $D_{\mathcal{H}}$ .
2. Set  $i = 0$  and  $D_0 = D_{\mathcal{H}}$ .
3. Choose  $S_i$  to be a non-empty set of vertices of  $D_i$  that have indegree zero. If there are no such vertices, then halt and return  $\mathcal{H}$  has no temporal representation.
4. Set  $D_{i+1} = D_i \setminus S_i$ . If  $D_{i+1}$  is the empty graph, then go to Step 5. Otherwise, increment  $i$  by 1 and go to Step 3.
5. Define  $f : V \rightarrow \mathbb{N}$  by setting  $f(v) = i$  for all  $v \in V$ , where  $[v] \in S_i$ .
6. Return the map  $f$ .

The correctness of this algorithm is guaranteed by the following result, whose proof is given in the Appendix.

**Theorem 5.2.** *Let  $\mathcal{H}$  be a hybrid, and suppose that TEMPREP is applied to  $\mathcal{H}$ .*

- (i) *If  $\mathcal{H}$  has a temporal representation, then TEMPREP returns a temporal labelling of  $\mathcal{H}$ .*
- (ii) *If  $\mathcal{H}$  has no temporal representation, then TEMPREP returns the statement  $\mathcal{H}$  has no temporal representation.*

Moreover, the running time of TEMPREP is polynomial in the size of the vertex set of  $\mathcal{H}$ .

For example, if one takes the hybrid phylogeny  $\mathcal{H}_1$  in Fig. 11(a) and apply the algorithm TEMPREP, we can reconstruct the temporal representation shown in

Fig. 9. Note that there is some choice as to the assignment of numbers for the leaves  $a$  and  $d$ . Such choices will generally arise for any hybrid phylogeny that has a temporal representation. Observe that it is the relative ordering of the vertices and not the actual values assigned by a temporal labelling that is important. We can make this idea more precise as follows.

Let  $\mathcal{H}$  be a hybrid with vertex set  $V$  that has a temporal representation, and let  $f_1$  and  $f_2$  be two temporal labellings of  $\mathcal{H}$ . We say that  $f_1$  and  $f_2$  are *ordering isomorphic* if, for all  $u, v \in V$ , the following hold:

- (i)  $f_1(u) < f_1(v)$  if and only if  $f_2(u) < f_2(v)$ ;
- (ii)  $f_1(u) = f_1(v)$  if and only if  $f_2(u) = f_2(v)$ .

Using the results in this section (and the Appendix) one can construct an algorithm that lists, up to ordering isomorphism, all temporal labellings of  $\mathcal{H}$  so that each such labelling is outputted in polynomial time. An outline of this algorithm is given in the Appendix. It is important to note that, as this list may be exponential in the size of  $V$ , the algorithm itself is not guaranteed to run in polynomial time.

## 6. ACKNOWLEDGEMENT

We thank Peter Lockhart for encouragement to develop some of the theory in this paper, and for helpful subsequent comments.

## REFERENCES

- [1] Bafna, V., and V. Bansal. 2004. The number of recombination events in a sample history: conflict graph and lower bounds. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 1(2):78–90.
- [2] Bang-Jensen, J., and G. Gutin. 2001. *Digraphs: Theory, Algorithms and Applications*. Springer-Verlag. London.
- [3] Baroni, M. 2004. Hybrid phylogenies: a graph-based approach to represent reticulate evolution. PhD thesis. University of Canterbury, Christchurch, New Zealand.
- [4] Baroni, M., C. Semple, and M. A. Steel. 2004. A framework for representing reticulate evolution. *Ann. Combin.* 8(4):391–408.
- [5] Baroni, M., S. Grünwald, V. Moulton, C. Semple. 2005. Bounding the number of hybridisation events for a consistent evolutionary history. *Journal of Mathematical Biology*. In press.
- [6] Bordewich, M., and C. Semple. 2004. On the computational complexity of the rooted subtree prune and regraft distance. *Ann. Combin.* 8(4):409–423.
- [7] Bordewich, M., and C. Semple. Computing the minimum number of hybridisation events for a consistent evolutionary history. Submitted.
- [8] FISHEER, F.J.F. 1965. *The alpine Ranunculi of New Zealand*. DSIR publishing, New Zealand.
- [9] Gusfield, D. 2004. A fundamental Decomposition Theory for Phylogenetic Networks and Incompatible Characters. Technical Report UC Davis CSE-2004-32.
- [10] Gusfield, D., S. Eddhu, C. Langley. 2004. Optimal, Efficient Reconstruction of Phylogenetic Networks with Constrained Recombination. *J. Bioinf. Comput. Biol.* 2(1):173–213 .
- [11] Holder, M. T., J. A. Anderson and A. K. Holloway. 2001. Difficulties in detecting hybridization. *Syst. Biol.* 50(6):978–982.
- [12] Holland, B., K. Huber, V. Moulton and P. J. Lockhart. 2004. Using consensus networks to visualize contradictory evidence for species phylogeny. *Mol. Biol. Evol.* 21:1459–1461.
- [13] Huson, D. H., T. Dezulian, T. Klopper, and M. A. Steel. 2004. Phylogenetic super-networks from partial trees. *IEEE Trans. Comput. Biol. Bioinf.* 1(4):151–158.
- [14] Huson, D. H., T. Klopper, P. J. Lockhart, M. A. Steel. 2005. Reconstruction of reticulate networks from gene trees. *Proceedings of RECOMB 2005*.
- [15] Legendre, P., and V. Makarenkov. 2002. Reconstruction of Biogeographic and Evolutionary Networks Using Reticulograms. *Syst. Biol.* 51(2):199–216.
- [16] Lockhart, P. J., P. A. McLenachan, D. Havell, D. Glenny, D. Huson, and U. Jensen. 2001. Phylogeny, radiation, and transoceanic dispersal of New Zealand alpine buttercups: molecular evidence under split decomposition. *Ann. Missouri Bot. Gard.* 88(3):458–477.
- [17] Maddison, W. P. 1997. Gene trees in species trees. *Syst. Biol.* 46(3):523–536.
- [18] Moret, B. M. E., *et al.* 2004. Phylogenetic Networks: Modeling, Reconstructibility, and Accuracy. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 1(1):1–11.
- [19] Song, Y., and J. Hein. 2003. On the minimum number of recombination events in the evolutionary history of DNA sequences. *J. Math. Biol.* 48:160–186.



## 7. APPENDIX

*Proof of Theorem 5.1*

A digraph  $D$  with vertex set  $V$  and arc set  $A$  has an *acyclic ordering* if there exists a map  $g : V \rightarrow \mathbb{N}$  such that, for all  $(u, v) \in A$ , we have  $g(u) < g(v)$ . The following lemma is well-known and easily proved (for example, see [2]).

**Lemma 7.1.** *A digraph is acyclic if and only if it has an acyclic ordering.*

**Proposition 7.2.** *Let  $\mathcal{H}$  be a hybrid with vertex set  $V$  and suppose that  $f : V \rightarrow \mathbb{N}$  is a temporal labelling of  $\mathcal{H}$ . Then  $\bar{f}$  induces an acyclic ordering of  $[V]$ . In particular,  $D_{\mathcal{H}}$  is acyclic.*

*Proof.* Let  $f : V \rightarrow \mathbb{N}$  be a temporal labelling of  $\mathcal{H}$ , and consider  $D_{\mathcal{H}}$ . Let  $([u], [v])$  be an arc of  $D_{\mathcal{H}}$ . To prove the proposition it suffices to show by Lemma 7.1 that  $\bar{f}([u]) < \bar{f}([v])$ . Now, by definition, there exists elements  $a \in [u]$  and  $b \in [v]$  such that  $(a, b)$  is a tree arc of  $\mathcal{H}$ . Since  $f$  is a temporal labelling of  $\mathcal{H}$ , we have that  $f(a) < f(b)$ , which in turn implies that  $\bar{f}([u]) < \bar{f}([v])$  as required.  $\square$

**Proposition 7.3.** *Let  $\mathcal{H}$  be a hybrid with vertex set  $V$ , and suppose that  $D_{\mathcal{H}}$  is acyclic. Let  $g$  be an acyclic ordering of  $[V]$ . Let  $f$  be the map from  $V$  into  $\mathbb{N}$  defined by setting  $f(v) = g([v])$ . Then  $f$  is a temporal labelling of  $\mathcal{H}$ .*

*Proof.* Let  $(u, v)$  be an arc of  $\mathcal{H}$ . First assume that  $(u, v)$  is tree arc. Then  $u$  and  $v$  are in different equivalence classes; otherwise,  $D_{\mathcal{H}}$  contains a loop contradicting the fact that  $D_{\mathcal{H}}$  is acyclic. Furthermore, there is an arc from  $[u]$  to  $[v]$  in  $D_{\mathcal{H}}$ . It now follows that  $f(u) < f(v)$ .

Now assume that  $(u, v)$  is a recombination arc of  $\mathcal{H}$ . Then  $[u] = [v]$ , and so  $f(u) = f(v)$ . Hence, by definition,  $f$  is a temporal labelling of  $\mathcal{H}$ .  $\square$

Combining Propositions 7.2 and 7.3, we obtain Theorem 5.1.  $\square$

*Proof of Theorem 5.2*

To see that TEMPREP does indeed work, we begin with the following well-known and easily proved lemma.

**Lemma 7.4.** *Let  $D$  be a digraph that contains no directed cycle. Then there exists a vertex of  $D$  whose indegree is zero.*

To prove part (i) of Theorem 5.2, suppose that  $\mathcal{H}$  has a temporal representation. Then, by Theorem 5.1,  $D_{\mathcal{H}}$  has no directed cycles. By Lemma 7.4, this implies that every subdigraph obtained from  $D_{\mathcal{H}}$  by deleting vertices (and their incident arcs) contains at least one vertex of indegree zero. It now follows that TEMPREP applied to  $\mathcal{H}$  returns a map  $f : V \rightarrow \mathbb{N}$ . To see that  $f$  is a temporal labelling of  $\mathcal{H}$ , define  $g : [V] \rightarrow \mathbb{N}$  by setting  $g([v]) = S_i$ , where  $[v] \in S_i$ . Because of the way in

which  $S_0, S_1, S_2, \dots$  are constructed,  $g$  is an acyclic ordering of the vertices of  $D_{\mathcal{H}}$ . Therefore, by Proposition 7.3, the map  $f$  is a temporal labelling of  $\mathcal{H}$ .

For the proof of part (ii) of Theorem 5.2 suppose that  $\mathcal{H}$  has no temporal representation. Then, by Theorem 5.1,  $D_{\mathcal{H}}$  contains a directed cycle. Let  $\{[v_1], [v_2], \dots, [v_k]\}$  be the vertices in this cycle, where we may assume that  $([v_j], [v_{j+1}])$  for all  $j$  and  $([v_k], [v_1])$  are arcs of this cycle. It is now easily seen that beginning with  $D_{\mathcal{H}}$ , and selecting and removing only vertices with indegree zero none of the vertices in this cycle can ever be removed. Thus at some iteration  $i$  of TEMPREP when applied to  $\mathcal{H}$ , no vertex of  $D_i$  has indegree zero, in which case TEMPREP halts and returns  $\mathcal{H}$  has no temporal representation. This completes the proof of (ii).

We leave the straightforward check that the running time of TEMPREP applied to  $\mathcal{H}$  is polynomial in the size of the vertex set of  $\mathcal{H}$  to the reader.

□

*Outline of an algorithm to output all temporal labellings of a hybrid phylogeny, up to order isomorphism*

By Proposition 7.2, all temporal labellings of  $\mathcal{H}$  induce an acyclic ordering of the vertex set  $[V]$  of  $D_{\mathcal{H}}$ . Conversely, by Proposition 7.3, all acyclic orderings of  $[V]$  induce a temporal labelling of  $\mathcal{H}$ . It follows that if  $\mathcal{H}$  has a temporal representation, then all temporal labellings of  $\mathcal{H}$  can be found by finding all acyclic orderings of  $[V]$ . Using the first part of the proof of Theorem 5.1, it is easily checked that all such orderings can be obtained by considering all possible ways of reducing  $D_{\mathcal{H}}$  to the empty graph by sequentially selecting and then deleting subsets of vertices of indegree zero. Since it is only the relative ordering of the vertices of  $\mathcal{H}$  that are of interest, it follows that it is only the order in which these subsets are chosen that are important. Each possible way of reducing  $D_{\mathcal{H}}$  to the empty graph gives rise to a unique sequence of chosen subsets of vertices of  $D_{\mathcal{H}}$ . In TEMPREP, this corresponds to all possible choices for the sequence  $S_0, S_1, S_2, \dots$ . Furthermore, each such sequence induces, up to ordering isomorphism, a unique temporal labelling of  $\mathcal{H}$ . Hence to list, up to ordering isomorphism, all temporal labellings of  $\mathcal{H}$  one simply needs to systematically find all possible choices for selecting  $S_0, S_1, S_2, \dots$  in TEMPREP.

□

BIOMATHEMATICS RESEARCH CENTRE, DEPARTMENT OF MATHEMATICS AND STATISTICS, UNIVERSITY OF CANTERBURY, CHRISTCHURCH, NEW ZEALAND

*E-mail address:* mbaroni@ugal.ro

*E-mail address:* c.semple@math.canterbury.ac.nz

*E-mail address:* m.steel@math.canterbury.ac.nz